The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.561
32(4), 561–572

Comparison of multiscale multiple change-points estimators

Jaehee Kim[a,1]

[a]Department of Statistics, Duksung Women's University

(Received April 16, 2019; Revised May 19, 2019; Accepted June 5, 2019)

Abstract

We study false discovery rate segmentation (FDRSeg) and simultaneous multiscale change-point estimator (SMUCE) methods for multiscale multiple change-point estimation, and compare empirical behavior via simulation. FSRSeg is based on the control of a false discovery rate while SMUCE used for the multiscale local likelihood ratio tests. FDRSeg seems to work best if the number of change-points is large; however, FDRSeg and SMUCE methods can both provide similar estimation results when there are only a small number of change-points. As a real data application, multiple change-points estimation is done with the well-log data.

Keywords: false discovery rate (FDR), FDRSeg, local likelihood ratio test, multiscale, multiple change-points, SMUCE

1. 서론

다중변화점에 의한 데이터 분할법은 데이터셋에 여러 개의 변화점이 존재하는 경우 통계적 추론에 있어 매우 중요한 문제이다. 다중변화점 추정문제는 금융데이터, 신호 데이터, 유전자공학 데이터 등 광범위한 데이터에 대해 적용이 필요한 문제이나 계산적으로 복잡하여 쉽지않은 문제이다. 모수적 분포를 아는 경우는 변화 검정에 우도비를 활용할 수 있으며 분포의 모수에 대한 변화점 검정통계량은 여러 연구자들에 의해 개발되었다. Chernoff와 Zacks (1964) 그리고 Kander와 Zacks (1966)를 시작으로 관심을 받았으며 Hinkley (1970), Hušková와 Antoch (2003), Siegmund (1988), Worsley (1983) 등이 우도비를 활용하여. 변화검정법 또는 한 개 변화점 추정법을 제안했다.

변화점의 개수를 모르는 경우는, 추가로 모형선택 단계가 필요하다. 이러한 경우에는 변화점의 개수와 관련하여 모형복잡도(model complexity)에 대한 적절한 벌점함수(penalty function)가 이용되어 벌점화 우도함수(penalized likelihood)를 최대화하는 방향으로 변화점을 찾게 된다. 여기서 벌점함수는 모수값과 변화점 개수에 대한 과적합(overfitting)을 피하도록 하는 역할을 한다. Yao (1988)와 Yao와

This research was supported by the Korea Research Foundation (KNRF) (No. 2018R1A2B26001664) as well as by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & Energy (MOTIE) of the Republic of Korea (No. 20161210200610).
[1]Department of Statistics, Duksung Women's University, 419 Samyang-ro 144 Gil 33, Dobong-Gu, Seoul 01369, Korea. E-mail: jaehee@duksung.ac.kr

Au (1989)는 BIC 기반 가중함수를 이용하였으며 Boysen 등 (2009)은 추가로 l_0-penalty를 이용한 방법을 제안하였다. 고차원인 경우 변화점 문제에 대해서는 Braun 등 (2000), Winkler와 Liebscher (2002)의 연구 결과가 있다. Zhang과 Siegmund (2007)는 변화점의 개수와 위치에 대한 벌점함수를 도입하여 변화점을 추정했다. 모형선택 기반 l_0-penalty 함수를 이용한 연구로는 변화회귀모형에 대한 연구로 Birge와 Massart (2006), Lavielle (2005), Lavielle와 Moulines (2000) 등이 있다. 벌점화 접근법으로 fused lasso 기법 (Friedman 등, 2007), Tibshirani 등 (2005), Harchaoui와 Levy-Leduc (2010)은 total variation과 l_1-norm penalty를 변화점 개수에 대한 convex surrogate로 놓았다. 베이지안기법 등을 이용한 다중변화점 추정에 대해서는 Cheon과 Kim (2010), Kim과 Cheon (2010, 2011), Kim과 Hart (2011) 등이 있다. 고차원 응용 다중변화점 문제로는 Levy-Leduc와 Roueff (2009)은 네트워크에 대한 다중변화점 추정법을 제안했다. 통계적 다층 제한(statistical multiscale constraint) 하에서 목적함수를 최소화하는 변화점 추정량 연구로는 Candes와 Tao (2007), Davies와 Kovac (2001), Davies 등 (2009), Frick 등 (2014)이 있다. Chan과 Walther (2013), Dümbgen과 Walther (2008)은 다층적 벌점화 방법 기반 통계량을 제안했다. 또한 Kolaczyk와 Nowak (2004), Zhang과 Siegmund (2012)는 다층적 분할법(multiscale partitioning)을 제안했고 Olshen 등 (2004)은 반복적 분할(recursive partitioning)을 제안했다.

본 연구에서는 다층적 다중변화점 추정법비교를 해보고자한다. 특히 동시다층변화점(simultaneous multiscale change-point estimator; SMUCE) (Frick 등, 2014)와 위발견률기분할기법(false discovery rate segmentation; FDRSeg) (Li 등, 2016) 방법에 의한 다중변화점 추정법의 각 특성과 방법을 설명하고 모의실험을 통해 상황별 비교를 하고자한다. 2장에서는 변화점 회귀모형(change-point regression model)에 대해 설명한다. 3장에서는 동시다층변화점추정법과 위발견률기반 다중변화점 추정법으로 Hybrid SMUCE 기법과 FDRSeg 기법을 설명한다. 4장에서는 여러 가지 다중변화점 상황에서 모의실험을 통해 다중변화점 추정 기법을 비교하고 실제 데이터 적용 예를 보인다. 마지막으로 5장에서는 간단한 결론을 맺는다.

2. 변화점 회귀모형

관측값에 대해 다음의 변화점 회귀모형을 고려한다.

$$Y_i = \mu\left(\frac{i}{n}\right) + \sigma\varepsilon_i, \quad i = 0, 1, \ldots, n-1, \tag{2.1}$$

여기서 오차항 $\epsilon_0, \ldots, \epsilon_{n-1}$은 $N(0,1)$을 따른다고 가정한다 ($\sigma > 0$). 평균함수 μ는 우연속 조각 상수 (right-continuous and piecewise constant) 함수로 $K+1$개 부분(segment)를 갖는다. 각 부분 $I_k = [\tau_k, \tau_{k+1}) \subset [0,1)$에 대해 $0 < \tau_1 < \cdots < \tau_k < \tau_{k+1} < \cdots < \tau_K < 1$이며

$$\mu = \sum_{k=0}^{K} c_k 1_{[\tau_k, \tau_{k+1})} \tag{2.2}$$

으로 표현하며 변화점 개수 K는 모르며(unknown) K개 변화점을 갖는다. k번째 부분 I_k에서는 c_k 값을 가지며 평균함수 μ의 식별성(identifiability)을 위해 $c_k \neq c_{k+1}$를 가정한다. 다중변화점 추정에 대한 일반적인 방법론은 변화점의 개수와 위치를 변경해보면서 벌점 비용함수(penalty cost function)를 최소화하도록 추정하는 것이다.

$$\min_{\mu} \sum_{k=0}^{K} C\left(Y_{[n\tau_k]}, \ldots, Y_{[n\tau_{k+1}]-1}; c_k\right) + \gamma_n f(K), \tag{2.3}$$

여기서 $C(\cdot)$는 비용함수, $f(K)$는 벌점함수로 고려하고 γ_n은 균형모수(balancing parameter)로 놓는다. 예를 들어, Boysen 등 (2009)은 희박성 부분집합 선택에 대한 벌점함수

$$f(K) = l_0(\mu) = K \tag{2.4}$$

를 고려하였다. 전체 최적화(global optimization) 방법의 대안으로는 반복적 계산에 의해 차례로 변화점을 추정해간다. 전체 최적화는 전체 구간에서의 비용함수등 해당 측도 기준으로 최적화 해를 구하는 것으로 모르는 개수의 다중변화점이 존재하는 경우 전체 최적화문제는 다루기 쉽지않다. 즉, 변화점이 추정되면 이분분할(binary segmentation)로 변화점 이전부분과 이후부분으로 나눈 후 다시 각 부분에서 각각 변화점을 추정하고 더 이상 변화점이 추정되지않을 경우 변화점 추정을 멈추고 다중변화점을 제공한다. 동적 프로그램(dynamic program) (Bellman, 1957; Bellman과 Dreyfus, 1962)을 이용한 다중 변화점 추정 연구가 지속되고 있으며 Olsen 등 (2004)은 circular binary segmentation (CBS) 기반, Fryzlewicz (2014)는 wild binary segmentation (WBS) 기반 다중변화점 추정법을 제안하였다. 그러나 이러한 이분 분할법은 국소오류는 조절가능하지만 전체오류율조절은 할 수 없는 단점이 있다.

3. 동시다층변화점추정법과 위발견률기반 다중변화점 추정법 비교

3.1. Hybrid SMUCE를 이용한 다중변화점 추정

Frick 등 (2014)은 혼합방법(hybrid method)으로 동시다층변화점 SMUCE 기법을 제안했는데 오류율 조정이 가능하며 변화점 개수를 최소화하는 방법으로 다층적 추정 구조를 포함한다. SMUCE 기법에서 j개의 가짜 변화점을 포함할 확률은

$$P\left(\hat{K} \geq K + j\right) \leq \alpha_S^{\left[\frac{j}{2}\right]}, \quad j = 1, 2, \ldots \tag{3.1}$$

으로 지수적 감소를 한다. 여기서 α_S는 미리 설정한 family-wise error rate (FWER)이다.

$$P\left(\hat{K} > K\right) \leq \alpha_S. \tag{3.2}$$

Frick 등 (2014)는 다음의 지수족 회귀모형(exponential family regression family model)

$$Y_i \sim F_{\vartheta\left(\frac{i}{n}\right)}, \quad i = 1, \ldots, n \tag{3.3}$$

을 고려한다. 여기서 $\{F_\theta\}_{\theta\in\Theta}$는 1차원 지수족에 포함하는 분포함수 집합이다. 확률밀도함수 f_θ는 K개 변화점을 갖는 우연속(right continuous) 함수이다.

$$\inf_{\vartheta\in S} \sharp J(\vartheta) \quad \text{subject to} \;\; T_n(Y, \vartheta) \leq q, \tag{3.4}$$

여기서 $\vartheta \in [0,1] \to \Theta \subset R$ 우연속 계단함수, $\sharp J(\vartheta)$는 변화점 즉 점프의 개수이고 q는 threshold이다. 이러한 문제를 풀기위해서는 최적화 문제(optimization problem)를 해결해야한다. Boysen 등 (2009)은 jump-penalized least squares regression 모형에서 piecewise constant function 추정 문제를 다루었다. Gaussian 경우를 고려하며 θ에 대한 다중다층 통계량(multiscale statistic)은 다음과 같이 정의한다.

$$T_n(Y, \theta) = \max_{\substack{1\leq i<j\leq n,\, \vartheta(t)=\theta \\ \text{for } t\in\left[\frac{i}{n}, \frac{j}{n}\right]}} \left[\sqrt{2T_i^j(Y,\theta)} - \sqrt{2\log\left(\frac{en}{j-i+1}\right)}\right]. \tag{3.5}$$

즉 $T_n(Y,\vartheta)$ 함수는 후보 변화함수 ϑ에 대한 multiscale statistic으로 정의한다. 여기서 $e = \exp(1)$, log는 자연로그(natural logarithm)이다. 구간 $[i/n, j/n]$에서 변화여부에 대한 검정으로 국소 우도비 통계량(local likelihood ratio statistic)은

$$T_i^j(Y,\theta_0) = \log\left\{\frac{\sup_{\theta\in\Theta}\prod_{l=i}^{j} f_\theta(Y_l)}{\prod_{l=i}^{j} f_{\theta_0}(Y_l)}\right\} \tag{3.6}$$

으로 정의된다. 이와 같이 구간별 검정을 통한 multiscale statistic T_n은 자연스럽게 다중다층 형태의 시스템에 대한 검정이 된다. 구간조정을 통한 검정으로부터 변화점에 대한 정보를 얻을 수 있다. 또한 $\#J$의 최소값이 변화점 개수 추정값이 된다. θ에 대한 추정량을 얻기위해 다음 집합

$$C(q) = \left\{\vartheta \in S : \sharp J(\vartheta) = \hat{K}(q) \quad \text{and} \quad T_n(Y,\vartheta) \le q\right\} \tag{3.7}$$

을 고려할 수 있고 신뢰집합(confidence set)을 구성한다. SMUCE $\hat{\vartheta}(q)$은 다음과 같이

$$\hat{\vartheta}(q) = \arg\max_{\vartheta\in C(q)} \sum_{i=1}^{n} \log\left\{f_{\vartheta\left(\frac{i}{n}\right)}(Y_i)\right\} \tag{3.8}$$

으로 $C(q)$ 내에서의 제한된 최대우도추정량으로 정의된다. 또한 다음의 확률부등식

$$P\left\{\hat{K}(q) > K\right\} \le \; P\left\{T_n(T,\vartheta) > q\right\} \tag{3.9}$$

을 얻을 수 있다. Frick 등 (2014)의 Theorem 2에서 변화점 개수를 과소추정할(probability of underestimating the number of change-points) 상한확률로

$$P\left\{\hat{K}(q) < K\right\} \le 2K\exp\left(-C_n\lambda\Delta^2\right)\left\{\exp\left(\frac{1}{2}\left[q + \sqrt{2\log\left(\frac{2e}{\lambda}\right)}\right]^2\right) + \exp\left(-3C_n\lambda\Delta^2\right)\right\} \tag{3.10}$$

를 보여준다. 여기서 $C > 0$는 분포족에 의존한 상수이고 확률값은 구간길이(interval length) λ, 최소 점프크기 Δ와 변화점 개수 K 에 의존한다. 비록 모든 스텝 함수 전체에서 uniform confidence set은 얻을 수 없지만 Δ와 λ가 작을수록 즉

$$\frac{n}{\log(n)}\Delta_n^2\lambda_n \to \infty \quad \text{as} \quad n \to \infty \tag{3.11}$$

일 때 근사적으로 uniform confidence set은 얻을 수 있다. 변화점 추정을 위한 dynamic programming은 local likelihood에 대해 제한을 두고 구할 수 있다. Killick 등 (2012)은 dynamic programming에 대한 여러가지 접근법을 언급하였다. Frick 등 (2014)의 Theorem 1에서 변화 통계량에 대한 극한 분포

$$T_n(Y,\vartheta;c_n) \to \max_{0\le k\le K}\sup_{\tau_k\le s<t\le\tau_{k+1}}\left[\frac{|B(t)-B(s)|}{\sqrt{(t-s)}} - \sqrt{2\log\left(\frac{e}{t-s}\right)}\right] \tag{3.12}$$

를 보여준다. 여기서 $B(t)$는 Brownian process이다.

변화점 개수 추정량은

$$\hat{K}(q) = \min\left\{k \in N, \exists\,\vartheta \in S_n[K] \,:\, T_n(Y,\vartheta;c_n) \le q\right\}, \quad q \in R \tag{3.13}$$

이다.

3.2. FDR 기반 다중변화점

Li 등 (2016)은 false discovery rate (FDR)을 기반으로 한 다중변화점 추정법으로 FDRSeg 기법을 제안하였다. FDR 조절을 통해 변화점 개수와 위치에 대한 추정능력을 높이고 특히 가우시안 상황에서는 FDR에 대한 상한(upper bound)를 계산하여 제공한다. SMUCE 방법에 비해 다층구조에서 FDR을 조절하면서 통계적 추정 효율성과 빠른 계산기법을 포함한다. 또한 다층 국소 우도검정(multiscale local likelihood test) 보다는 다중검정문제해결로 FDR을 조절하고자한다. 이러한 조절을 위해서는 Benjamini와 Hochberg (1995)의 FDR 계산을 활용한다. $\hat{\tau}_i$이 다음의 구간

$$\left[\frac{\lceil n(\hat{\tau}_{i-1}+\hat{\tau}_i)/2\rceil}{n}, \frac{\lceil n(\hat{\tau}_i+\hat{\tau}_{i+1})/2\rceil}{n}\right] \tag{3.14}$$

에 속하면 참발견(true discovery)으로 분류한다. $\hat{\tau}_0 = 0$, $\hat{\tau}_{K+1} = 1$, FDR은 다음과 같이 정의한다.

$$\text{FDR} = E\left[\frac{\text{FD}}{\hat{K}+1}\right], \tag{3.15}$$

여기서 FD는 거짓발견(false discovery)된 변화점 개수이다. 평가측도로서

$$d(\mu, \hat{\mu}) = \max_{0\le i\le K+1} \min_{0\le j\le \hat{K}+1} |\tau_i - \hat{\tau}_j| \tag{3.16}$$

를 고려한다. 여기서

$$\mu = \sum_{i=0}^{K} 1_{[\tau_i, \tau_{i+1})} c_i,$$
$$\hat{\mu} = \sum_{j=0}^{\hat{K}} I_{[\hat{\tau}_j, \hat{\tau}_{j+1})} \hat{c}_j$$

이다. Li 등 (2016)의 FDR조절 방법을 살펴보자.

$$T_I(Y, c) = \max_{[\frac{i}{n}, \frac{j}{n}]\subset I} \frac{\left|\sum_{l=i}^{j}(Y_l - c)\right|}{\sigma\sqrt{j-i+1}} - \text{pen}\left(\frac{j-i+1}{\sharp I}\right), \tag{3.17}$$

여기서 c는 실수(real number), $\text{pen}(x) = \sqrt{2\log(e/x)}$는 scale에 대한 벌점항이고 $\sharp I$는 sampling point 개수이다. $\alpha \in (0,1)$에 대해 local quantile $q_\alpha(m)$은 다음과 같이 정의한다.

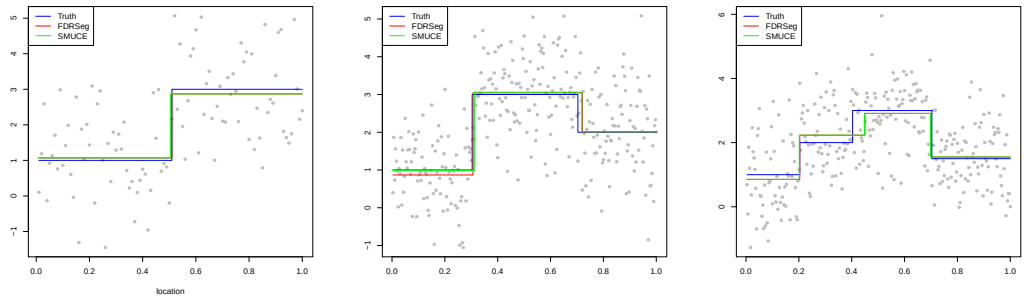
$$q_\alpha(m) = \min\left\{q : P\left(T_I(\varepsilon, \bar{\varepsilon}_I) > q\right) \le \alpha\right\}, \tag{3.18}$$

여기서 $\varepsilon = (\varepsilon_0, \ldots, \varepsilon_{n-1})$은 표준정규분포를 따르는 확률변수이며

$$\bar{\varepsilon}_I = \frac{1}{(\sharp I)} \sum_{\frac{i}{n}\in I} \varepsilon_i \tag{3.19}$$

이다. 다층 조건은 다음과 같고

$$C_k = \left\{\mu = \sum_{i=0}^{k} c_i 1_{I_i} \ : \ T_{I_i}(Y, c_i) - q_\alpha(I_i) \le 0, \quad \text{for all} \ \ i = 0, 1, \ldots, k\right\} \tag{3.20}$$



(a) (b) (c)

Figure 4.1. (a) One change-point case; (b) Two change-points case; (c) Three change-points case. FDRSeg = false discovery rate segmentation; SMUCE = simultaneous multiscale change-point estimator.

변화점의 개수는

$$\hat{K} = \min\{k : C_k \neq \emptyset\} \tag{3.21}$$

로 추정한다. FDRSeg 추정량은

$$\hat{\mu} = \arg\min_{\mu \in C_{\hat{K}}} \sum_{i=0}^{n-1} \left(Y_i - \mu\left(\frac{i}{n}\right)\right)^2 \tag{3.22}$$

으로 추정한다. 즉 $C_{\hat{K}}$ 내부의 추정량이 된다.

4. 모의실험 및 실제 데이터 분석

4.1. 모의실험

FDRseg와 SMUCE 변화점 추정능력을 비교하기 위해 여러 형태의 변화점이 발생하는 상황에서의 데이터에 대해 모의실험을 하고자한다. 오차항 ϵ_i는 서로 독립적이며 평균 0, 분산 σ^2인 정규분포를 따른다. 각 경우에 대한 표본 크기 n이고 실험 반복수는 1,000번이다. 추정량의 추정능력 측도로는 평균 변화점개수의 평균을 계산하고 평가측도로서 FDR 식 (3.15)와 $d(\mu, \hat{\mu})$ 식 (3.16)을 계산한다. 추정된 변화점 개수 $\hat{K}$에 대해 추정된 변화점을 $\{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\}$로 놓는다. $\hat{\tau}_0 = 0$, $\hat{\tau}_{\hat{K}+1} = 1$, $t = i/n$, $t \in [0, 1]$로 놓는다.

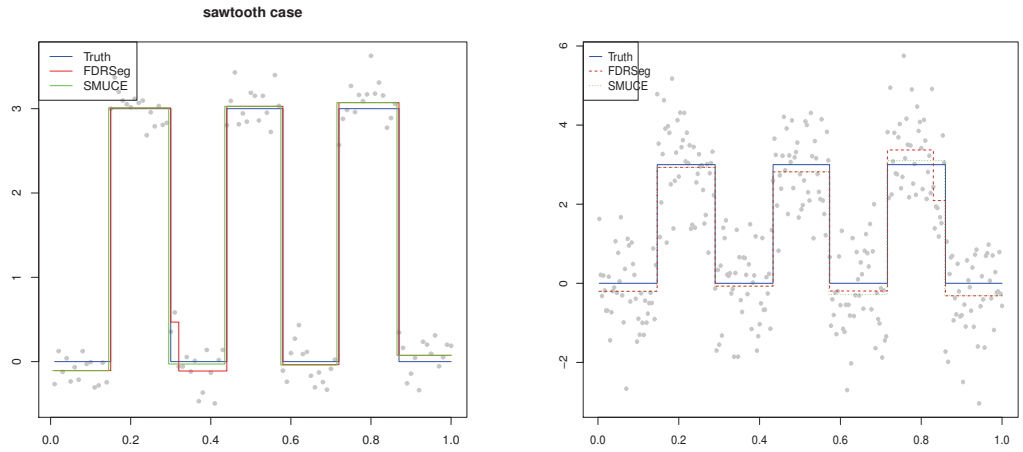
(i) 변화점 1개인 경우, $n = 100$

$$y_i = \begin{cases} 1 + \epsilon_i, & t < \tau, \\ 3 + \epsilon_i, & t \geq \tau. \end{cases}$$

$\tau = 0.50$인 경우 FDRseg와 SMUCE 모두 정확히 τ를 추정하고 변화점으로 나누어진 각 부분에 대한 평균함수를 추정한다 (Figure 4.1(a)).

(ii) 변화점 2개인 경우, $n = 300$

$$y_i = \begin{cases} 1 + \epsilon_i, & t < \tau_1, \\ 3 + \epsilon_i, & \tau_1 \leq t < \tau_2, \\ 2 + \epsilon_i, & t \geq \tau_2. \end{cases}$$



(a) (b)

Figure 4.2. (a) Sawtooth function case with $\sigma = 0.5$; (b) Sawtooth function case with $\sigma = 1$. FDRSeg = false discovery rate segmentation; SMUCE = simultaneous multiscale change-point estimator.

$\tau_1 = 0.3$, $\tau_2 = 0.7$인 경우 FDRseg와 SMUCE 모두 정확히 τ_1과 τ_2를 추정하고 변화점으로 나누어진 각 부분에 대한 평균함수를 추정한다 (Figure 4.1(b)).

(iii) 변화점 3개인 경우, $n = 300$

$$y_i = \begin{cases} 1 + \epsilon_i, & t < \tau_1, \\ 2 + \epsilon_i, & \tau_1 \le t < \tau_2, \\ 3 + \epsilon_i, & \tau_2 \le t < \tau_3, \\ 1.5 + \epsilon_i, & t \ge \tau_3. \end{cases}$$

$\tau_1 = 0.2$, $\tau_2 = 0.5$, $\tau_3 = 0.7$인 경우 FDRseg와 SMUCE 변화점 추정결과를 보여주고 변화점으로 나누어진 각 부분에 대한 평균함수를 추정한다 (Figure 4.1(c)).

(iv) 톱니모양 함수이며 변화점 6개인 경우, $n = 300$ (Figure 4.2).

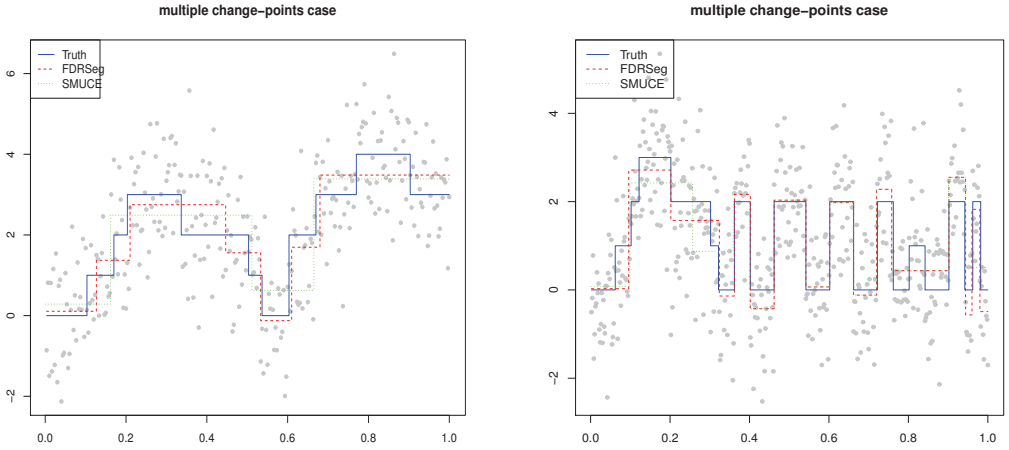
(v) 계단함수이며 변화점 10개인 경우, $n = 300$ (Figure 4.3(a)).

(vi) 계단함수이며 변화점 20개인 경우, $n = 500$ (Figure 4.3(b)).

Figure 4.2는 변화점 6개 가진 톱니 함수의 변화함수를 가지며 오차항의 표준편차가 이 $\sigma = 0.5$인 경우 추정결과이다. Figure 4.3(a)와 (b)는 여러 개 변화점을 가진 계단함수에서 추정 결과 예를 보여준다. Table 4.1에서는 여러 경우에서의 두 기법간의 변화점 추정능력 비교를 보여주며 변화점 개수가 많을수록 두 방법간 변화점 추정에 차이가 남을 볼 수 있다.

4.2. 검층 주상도 데이터 에 대한 다중변화점 추정

변화점 추정방법을 실제 데이터 검층 주상도 데이터(well-log data)에 적용하고 비교해보고자 한다. 이 데이터는 유전을 뚫는 동안 얻어진 지하 암반의 핵 자기적 반응으로 총 4,050개 관측값으로 구성된다. Figure 4.4에서 well-log 데이터를 보여준다. 이 데이터는 유전을 둘러싸고 있는 암석의 물리



(a) (b)

Figure 4.3. (a) multiple change-points $K = 10$; (b) multiple change-points $K = 20$. FDRSeg = false discovery rate segmentation; SMUCE = simultaneous multiscale change-point estimator.

Table 4.1. Comparison of FDRseg and SMUCE with $\alpha = 0.05$

	Case	Method	No. change-points	$d(\mu, \hat{\mu})$	FDR
(i)	$\tau = 0.5$	FDRseg	1.186	0.663	0.048
		SMUCE	1.051	0.657	0.017
(ii)	$\tau_1 = 0.3$	FDRseg	2.160	3.804	0.035
	$\tau_2 = 0.7$	SMUCE	2.011	3.732	0.0035
(iii)	$\tau_1 = 0.2$	FDRseg	3.070	8.917	0.0245
	$\tau_2 = 0.4, \tau_3 = 0.7$	SMUCE	2.660	19.636	0.0014
(iv)	Sawtooth $\sigma = 0.5$	FDRseg	6.247	1.241	0.028
	$\tau_1 = 0.2$, $\tau_2 = 0.4, \tau_3 = 0.7$	SMUCE	6.002	1.200	0.00025
(v)	10 change-points	FDRseg	6.979	28.466	0.0195
		SMUCE	4.641	50.937	0.00026
(vi)	20 change-points	FDRseg	16.637	24.363	0.0192
		SMUCE	9.941	46.706	0.0009

FDRSeg = false discovery rate segmentation; SMUCE = simultaneous multiscale change-point estimator.

적 구조를 이해하는데 사용되며, 평균에서의 변화들은 지각의 층화를 반영한다. 신호들은 거의 조각 상수(piecewise constant)이고, 각 상수 부분은 일정한 물리적 성질들을 가진 단일 암석층과 관련되어있다. 반면에 신호가 불연속적인 부분들은 새로운 암석층을 만날 때 발생한다. 즉 암석층의 변화는 신호가 불연속적일 때의 변화점들과 일치한다. 이러한 변화점을 발견하여 암석의 종류에 따른 변화를 예측하는 것은 석유 시추에 있어 중요한 문제이다. 암석에서 유동체들에 의해 가해지는 압력이 시추공에서의 압력을 초과할 때 굴착유체나 석유, 물이 갑자기 조절할 수 없을 만큼 분출되는 '폭발'이 발생하는데 이러한 문제가 새로운 암석층이 만날 때마다 시추공에서의 압력을 조절하는 것으로 방지될 수 있기 때문이다. 따라서 well-log 데이터의 신호에서 불연속점들을 발견하는 문제는 실제적으로 중요한 문제이다.

이 데이터에 대한 변화점 추정을 한 기존 연구를 살펴보고자 한다. Fearnhead (2006)는 베이지안 다중변화점 기법으로 추정시 40개 변화점을 추정하였고 Kim과 Cheon (2013)은 베이지안 기법 기반으로

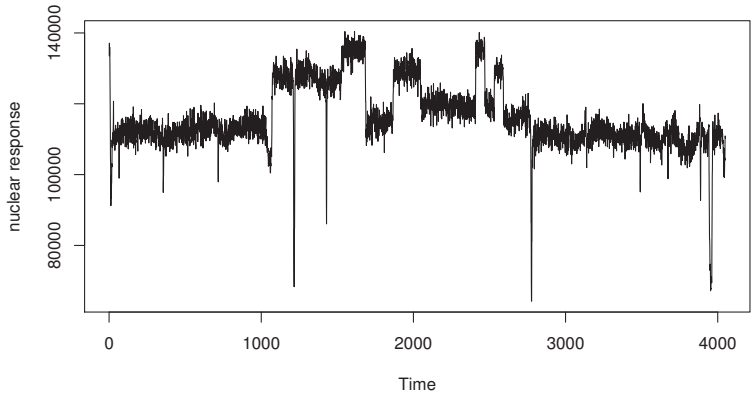


Figure 4.4. Well-log data.

Table 4.2. Comparison of change-points estimation with well-log data

Method	No. of change-points	
Fearnhead (2006)	40	
Kim and Cheon (2013)	19	Bayesian
SMUCE (2014)	49	
FDRseg (2016)	89	FDR control

SMUCE = simultaneous multiscale change-point estimator; FDR = false discovery rate; FDRSeg = FDR segmentation.

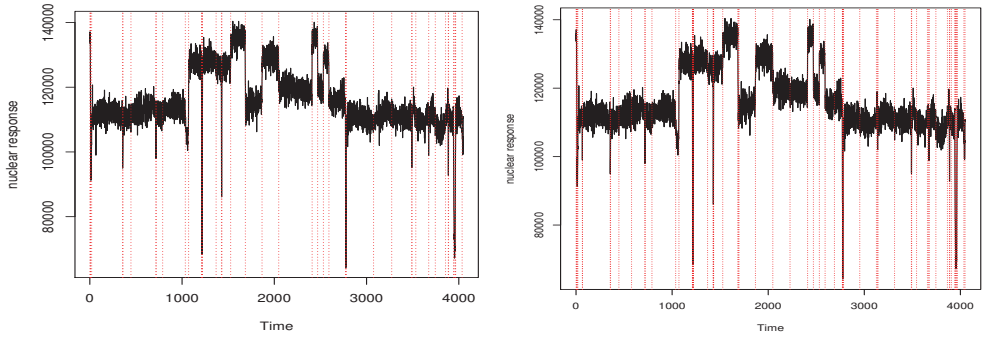


Figure 4.5. (a) SMUCE change-points estimation; (b) FDRSeg change-points estimation. SMUCE = simultaneous multiscale change-point estimator; FDRSeg = false discovery rate segmentation.

다변량 정규분포를 활용해 19개의 변화점을 추정하였다. SMUCE 기법으로는 49개 FDRSeg 기법으로는 89개 변화점이 추정되어 추정방법에 따른 추정된 변화점 개수가 차이가 큰 편이다. Table 4.2에서는 well-log 데이터에 대한 기존 변화점 추정 결과를 비교하는데 FDRSeg에 의해서 가장 많은 수의 변화점을 추정한다. Table 4.2를 보면 실제 데이터분석 결과 변화점 추정 개수가 차이가 많이 난다. 변화량 등을 포함한 변화점 추정법 기준에 따라 개수가 달라질 수 있다. FDRSeg 방법이 local change 에 따른 변화점을 더 많이 감지하는 것으로 생각되나 좀 더 연구가 필요해 보인다. Figure 4.5에서는 SMUCE 와 FDRSeg 기법으로 변화점 추정후 세로선으로 변화점 위치를 보여준다.

5. 결론

최근 빅데이터는 여러 분야에서 발생하며 빅데이터 분석 문제에서 작은 구간 또는 작은 그룹의 변화도 관심이 있을 수 있다. 이러한 면에서 다중변화점 문제는 실제 해결해야할 문제이며 벌점기반 최적화 문제를 포함하므로 계산 알고리즘과 더불어 해결해야할 과제이다. SMUCE 기법은 지수족 데이터에 대해 국소우도비검정을 이용한 다중변화점 추정방법으로 벌점기반 최적화 문제에서 제한된 최적화 문제이며 벌점 비용 함수로 다시 표현할 수 있으며 계층구조 조절을 통한 추정이 가능하다. FDR 기법은 FWER 또는 FDR 기반 방법으로 의미있는 해석이 가능하며 독립적/비독립적 데이터에 대해 적용가능하다. 본 연구에서는 SMUCE 기법과 FDRSeg 기법에 대해 이론적 특징과 모의실험을 통한 경험적 특성을 보였다. 변화점의 개수가 많을수록 FDRSeg 기법의 추정이 우수했는데 FDR 조절을 통해 거짓 점프의 개수가 제한되도록 한 이유도 있어보인다. 데이터의 상황을 반영하여 적절한 다중변화점 추정법 선택을 통한 추정이 가능하며 이와 같은 동시 다중변화점 추정방법의 활용과 연구가 기대된다.

References

Bellman, R. (1957). *Dynamic Programming*, Princeton University Press, Princeton, NJ.

Bellman, R. E. and Dreyfus, S. E. (1962). *Applied Dynamic Programming*, Princeton University Press, Princeton, NJ.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *Journal of Royal Statistical Society. Series B (Methodological)*, **57**, 289–300.

Birgé, L. and Massart, P. (2006). Minimal penalties for Gaussian model selection, *Probability Theory and Related Fields*, **138**, 33–73.

Boysen, L., Kempe, A., Liebscher, V., Munk, A., and Wittich, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators, *The Annals of Statistics*, **37**, 157–183.

Braun, J. V., Braun, R. K., and Müller, H. G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation, *Biometrika*, **87**, 301–314.

Candes, E. and Tao, T. (2007) The Dantzig selector: statistical estimation when p is much larger than n, *The Annals of Statistics*, **35**, 2313–2351.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio, *Statistica Sinica*, **23**, 409–428.

Cheon, S. and Kim, J. (2010). Multiple change-point detection of multivariate mean vectors with Bayesian approach, *Computational Statistics & Data Analysis*, **54**, 406–425.

Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to change in time, *The Annals of Mathematical Statistics*, **35**, 999–1018.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution, *The Annals of Statistics*, **29**, 1–65.

Davies, P. L., Kovac, A., and Meise, M. (2009). Nonparametric regression, confidence regions and regularization, *The Annals of Statistics*, **37**, 2597–2625.

Dümbgen, L. and Walther, G. (2008). Multiscale inference about a density, *The Annals of Statistics*, **36**, 1758–1785.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems, *Statistics and Computing*, **16**, 203–213.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change-point inference, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, with discussion and rejoinder by the authors, **76**, 495–580.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization, *Annals of Applied Statistics*, **1**, 302–332.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection, *The Annals of Statistics*, **42**, 2243–2281.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty, *Journal of the American Statistical Association*, **105**, 1480–1493.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables, *Biometrika*, **57**, 1–17.

Hušková, M. and Antoch, J. (2003). Detection of structural changes in regression, *Tatra Mountains Mathematical Publications*, **26**, 201–215.

Kander, Z. and Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time points, *The Annals of Mathematical Statistics*, **37**, 1196–1210.

Kim, J. and Cheon, S. (2010). A Bayesian regime-switching time-series model, *Journal of Time Series Analysis*, **31**, 365–378.

Kim, J. and Cheon, S. (2011). Bayesian multiple change-point estimation with annealing stochastic approximation Monte Carlo, *Computational Statistics*, **25**, 215–239.

Kim, J. and Hart, J. D. (2011). A change-point estimator using local Fourier series, *Journal of Nonparametric Statistics*, **23**, 83–98.

Kim, J. H. and Cheon, S. Y. (2013). Bayesian multiple change-point estimation and segmentation, *Communications for Statistical Applications and Methods*, **20**, 439–454.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost, *Journal of the American Statistical Association*, **107**, 1590–1598.

Kolaczyk, E. D. and Nowark, R. D. (2004). Multiscale likelihood analysis and complexity penalized estimation, *Annals of Statistics*, **32**, 500–527.

Lavielle, M. (2005). Using penalized contrasts for the change-point problem, *Signal Processing*, **85**, 1501–1510.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series, *Journal of Time Series Analysis*, **21**, 33–59.

Lévy-Leduc, C. and Roueff, F. (2009). Detection and localization of change-points in high-dimensional network traffic data, *Annals of Applied Statistics*, **3**, 637–662.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation, *Electronic Journal of Statistics*, **10**, 918–959.

Olshen, A. B., Venkatraman, E. S., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data, *Biostatistics*, **5**, 557–572.

Siegmund, D. (1988). Confidence sets in change-point problems, *International Statistical Review / Revue Internationale de Statistique*, **56**, 31–48.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused LASSO, *Journal of the Royal Statistical Society Series B (Statistical Methodology)*, **67**, 91–108.

Winkler, G. and Liebscher, V. (2002). Smoothers for discontinuous signals, *Journal of Nonparametric Statistics*, **14**, 203–222.

Worsley, K. J. (1983). The power of likelihood ratio and cumulative sum tests for a change in a binomial probability, *Biometrika*, **70**, 455–464.

Yao, Y. C. (1988). Estimating the number of change-points via Schwarz criterion, *Statistics & Probability Letters*, **6**, 181–189.

Yao, Y. C. and Au, S. T. (1989). Least-squares estimation of a step function, *Sankhyā: The Indian Journal of Statistics, Series A*, **51**, 370–381.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data, *Biometrics*, **63**, 22–32.

Zhang, N. R. and Siegmund, D. O. (2012). Model selection for high-dimensional, multi-sequence change-point problems, *Statistica Sinica*, **22**, 1507–1538.

SMUCE와 FDR segmentation 방법에 의한 다중변화점 추정법 비교

김재희[a,1]

[a]덕성여자대학교 정보통계학과

(2019년 4월 16일 접수, 2019년 5월 19일 수정, 2019년 6월 5일 채택)

요 약

본 연구는 다층적 다중변화점 추정법으로 FDRSeg 기법과 SMUCE 기법의 이론적 특성을 파악하고 모의실험을 통해 경험적 특성을 비교하고자한다. FDRSeg (False discovery rate segmentation)기법은 FDR 기반 조절을 하여 변화점을 추정하고 SMUCE (simultaneous multiscale change-point estimator) 기법은 국소우도함수 기반 다중검정으로 변화점을 추정한다. 변화점의 개수가 작을경우에는 두 기법에 의한 추정능력이 비슷하다. 변화점 개수가 많을수록 FDRSeg 의 추정이 변화점 개수와 추정측도 면에서 더 좋은 편이다. 실제 데이터 분석으로 검층 주상도 데이터에 대해 각 기법으로 다중변화점 추정을 하고 비교한다.

주요용어: 위발견율, FDRSeg, 국소우도함수, 다층적, 다중변화점, SMUCE

이 논문은 한국연구재단(KNRF)의 지원을 받아 수행된 연구과제입니다 (No. 2018R1A2B26001664). 또한 본 연구는 한국전력공사의 2018년 착수 에너지 거점대학 클러스터 사업에 의해 지원되었습니다 (Grant number: R18XA01).

[1](01369) 서울시 도봉구 삼양로 144길 33, 덕성여자대학교 정보통계학과. E-mail: jaehee@duksung.ac.kr